Acta Analytica (2021) 36:191–212
https://doi.org/10.1007/s12136-020-00447-9

Extending the Transparency Method beyond Belief: a Solution to the Generality Problem

Adam J. Andreotta[1]

Received: 14 February 2020 / Accepted: 27 July 2020/ Published online: 10 August 2020

Abstract

According to the Transparency Method (TM), one can know whether one believes that *P* by attending to a question about the world—namely, 'Is *P* true?' On this view, one can know, for instance, whether one believes that Socrates was a Greek philosopher by attending to the question 'Was Socrates a Greek philosopher?' While many think that TM can account for the self-knowledge we can have of such a belief—and belief in general—fewer think that TM can be generalised to account for the self-knowledge we can have of other propositional attitudes, such as our desires, intentions, wishes and so on. Call this the *Generality Problem*. In the present paper, I contrast my own attempt to solve the Generality Problem with several recent ones. I argue that in order to extend TM beyond belief, we must look to the concepts underpinning each kind of mental state. Doing so, I argue, reveals a series of outward-directed questions that can be attended to, in order to know what one desires, intends, wishes and so on. Call this the *conceptual approach* to extending TM. I support the conceptual approach in the present paper by showing how it generates Moore-Paradoxical sentences that are analogous to the case of belief.

Keywords Self-knowledge · Transparency · Introspection · Belief · Desire · Intention

✉ Adam J. Andreotta
adamandreotta@outlook.com

[1] Philosophy Department, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia

1 Introduction

In the past few years, a growing dissatisfaction with the idea that we can know our own mental states by an inward glance, or *inner sense*, has led to the development of several rival accounts of self-knowledge.[1] One of the most widely discussed of these is the transparency method (hereafter, 'TM').[2] On this view, one can know whether one currently has a mental state by attending to, in the requisite manner, a corresponding question about the world. For example, one can know whether one currently believes that Socrates was a Greek philosopher by answering the worldly question 'Was Socrates a Greek philosopher?'

TM has been defended in various forms by Evans (1982), Gordon (2007), Moran (2001, 2004, 2012), Byrne (2005, 2011, 2012, 2018), Bilgrami (2006), Boyle (2011, 2019) and Fernández (2013), among others. And it has been criticised by Nichols and Stich (2003), Bar-On (2004), Gertler (2011), Carruthers (2011) and Cassam (2014), among others. One of the most common objections that these critics advance is that TM is limited in its applicability. While they grant that TM may be able to yield self-knowledge in *some* circumstances—such as with one's current beliefs about the nationality of ancient Greek philosophers—they argue that TM cannot account for more complex belief states, and other types of propositional attitudes such as one's desires, intentions and wishes. They conclude, therefore, that TM is not a genuine competitor to the inner sense view which does purport to give a complete account of how we gain self-knowledge of these states. Following Cassam (2014, p. 103), call this the *Generality Problem*.[3]

In the present paper, I propose a general strategy for solving the Generality Problem. I argue that there are instances of each category of propositional attitude (e.g. desires, intentions, wishes and so on) where TM is applicable.[4] My contention is that in order to extend TM beyond belief, we must look to the concepts underpinning each kind of propositional attitude. Doing so, I argue, reveals a series of outward-directed questions that can be attended to, in order to know what one desires, intends, wishes and so on. Call this the *conceptual approach* to extending TM.

I proceed as follows. In Section 2, I explicate TM. In doing so, I show how the application of TM to belief can be supported by appealing to Moore's Paradox. In Section 3, I present the Generality Problem for TM. In Section 4, I critique some recent attempts to extend TM to desire—namely, those of Fernández (2013) and Byrne (2018). I then present the conceptual approach, which I argue does a better job of explaining self-knowledge of desire. In Section 5, I critique some recent attempts to

[1] Proponents of the inner sense view include Lycan (1996), Armstrong (1968), Nichols and Stich (2003) and Goldman (2006). According to this view, we acquire self-knowledge of our mental states by *looking inside*, or by employing our own 'internal monitor', as Lycan (1996, p. 33) puts it. For a discussion of some of the most recent objections to the inner sense view, see Byrne (2018, Ch. 2). While Byrne does not think that any of these objections present a 'knock-down refutation' (2018, p. 49) of the view, he does think that there are 'grounds for dissatisfaction [and that it] is time to examine some leading alternatives' (2018, p. 49).

[2] Another rival view to the inner sense view is called *neo-expressivism*. It is defended by Bar-On (2004) and Finkelstein (2003). According to this view, we do not *detect* our mental states, but rather *express* them. I will not consider this view in this paper.

[3] Gordon, alternatively, calls this the 'belief only' (2007, p.155) objection.

[4] Like Byrne (2018), I also think that TM can be extended to sensations. In this paper, however, I am only concerned with showing how TM can be applied to propositional attitudes.

extend TM to intention by Byrne (2018) and Paul (2012). I then defend the conceptual approach. In Section 6, I do the same for wishes—first critiquing Barz's (2015) recent attempt to extend TM to wishes before defending the conceptual approach.

2 TM and Belief

According to TM, one can know what one currently believes (something psychological) by attending to a question about the world (something non-psychological). The *locus classicus* is from Gareth Evans, who says the following:

> In making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?", I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" (1982, p. 225).

Here, Evans articulates what I take to be TM's positive thesis—namely, that we gain knowledge of our mental states by attending to the features of the world that our mental states are about. Another way of putting this is to say that we answer an 'inward-directed' question by attending to an 'outward-directed' question.[5] The process is known as 'transparent' because one 'looks through' the question 'Do I believe that *P*?' to focus on what the belief is about (the intentional object of the belief).

TM, as I interpret it, also involves a negative thesis. Evans mentions this when he says, 'I shall quite avoid the idea of this kind of self-knowledge as a form of perception' (1982, p. 225); and, 'in order to understand the self-ascription of experience we need to postulate no special faculty of inner sense or internal self-scanning' (1982, p. 230). The negative thesis that Evans attributes to TM suggests that we do not learn about our mental states by (1) looking inward and detecting the presence of a mental state, or by (2) observing our behaviour.[6] This is not to say that TM rules out the possibility that we can learn about our minds by observing our behaviour—such as when someone comes to believe that they are angry when they see their own refection in the mirror. It is only to say that this would not count as following TM (see Ashwell 2013, p. 247).

Since my primary aim in the present paper is to show how TM can be extended beyond belief, I will not provide a thoroughgoing defence of the claim that TM can account for the self-knowledge we can have of our beliefs.[7] It is, however, important

[5] In contrasting 'inward-directed' questions with 'outward-directed' questions, I am following Cassam (2014, p. 3), Moran (2004, p. 457) and Finkelstein (2003, p. 161).

[6] Moran also attributes this negative thesis to TM when he says that the view does not involve 'an "inward" glance or… observation of one's own behaviour' (2001, p. 101). Not all philosophers attribute this negative thesis to TM, however. Silins (2012), for example, says, 'it remains perfectly possible that we obtain self-knowledge through inner observation and the transparency method at the same time' (2012, p. 305). In my view, Silins' 'compatibilism', which entails a rejection of the negative thesis, is in tension with TM's positive claim. This is a point I do not have the space to argue for in this paper, however.

[7] For an account of how TM is applicable to belief, see Moran (2001), Fernández (2013) and Byrne (2005, 2018).

that I outline one of the key reasons for accepting the view—namely Moore's Paradox (hereafter, 'MP'). This is because MP will feature in my attempt to extend TM beyond belief.[8]

In thinking that MP can tell us something informative about the way in which we know our beliefs, I am following Shoemaker (1996, Ch. 4), Moran (2001) and Silins (2012). I agree with Silins, for example, who says that MP supports the 'transparency thesis' (2012, p. 302)—the thesis that one can know whether one believes that *P* by determining whether they judge that *P* is the case (this being the positive thesis we attributed to TM earlier). I will argue below that MP can also tell us something important about the self-knowledge we can have of our non-doxastic propositional attitudes.

Before turning to the question of how to extend TM beyond belief, let us first get clear about what MP is. We can do so by looking at conjunctive sentences of the following form: '*P*, but I don't believe that *P*'; 'I believe that *P*, but it is not the case that *P*'; and '*P* is true, but I don't believe *P* is true'.[9] For example, it would be absurd for a native English speaker to utter or imagine the following sentence:

[1] Nouméa is the capital city of New Caledonia, but I do not believe that it is.

The puzzling thing about [1], as Moore (1942) famously pointed out, is that while a speaker who uttered it would be speaking absurdly, she need not be uttering anything contradictory. While there is some controversy about why this is so, I think, along with Moore—and Sydney Shoemaker (1996, pp. 74–77)—that this is because while what the speaker *says* in [1] is not a contradiction, what she *implies* is. What the speaker says in [1] could be true because the sentence consists of two independent claims: one is a factual claim about the world—namely, that Nouméa is the capital city of New Caledonia; and the second is a psychological self-ascription—namely, the speaker's belief about a capital city of a French territory. The paradox arises in [1] because of what the speaker *implies*. This is because when the speaker utters the first conjunct of the sentence, she seems to affirm a statement about the world—that is, she seems to be judging a certain state of affairs to be the case. It is for this reason that it is tempting to attribute the following to our speaker:

[2] I believe that Nouméa is the capital city of New Caledonia, but I do not believe it is.

Now clearly [2] involves a contradiction. This because it is stated explicitly in terms of the subject's beliefs. This is unlike [1], which involves a judgement and a belief. What I think this shows is that when one judges that *P* is the case, one will also believe that *P*. In my view, we are entitled to think this is the case because MP shows the implausibility of accepting the

[8] Another reason to accept TM, some have argued, is that is captures the phenomenology of self-attributing a belief. See Valaris (2014) for a defence of this claim.

[9] Moore's Paradox is named after the English philosopher G.E. Moore. As Green and Williams (2007, p. 5) point out, Moore is careful to distinguish between paradox and absurdity. Moore claims it would be absurd for a speaker to utter a MP sentence; and it is paradoxical that such an absurd sentence is not contradictory. The following MP sentence is an example that Moore himself provides: 'I went to the pictures last Tuesday, but I don't believe I did' (1942, p. 543).

falsehood of the claim—namely, that one might judge that *P* is the case and yet not believe that *P*.

As is also common to point out, MP only seems to arise from the first-person point of view. There would be no absurdity associated with the following sentence (provided I am not Bart):

[3] Nouméa is the capital city of New Caledonia, but Bart does not believe it is.

And neither is it absurd in a case involving my former self:

[4] Nouméa is the capital city of New Caledonia, but I did not believe it was last week.

There is nothing absurd about either sentence: [3] is perfectly intelligible because Bart may have never heard of New Caledonia, and so he does not know what its capital city is, even if I judge that it is Nouméa; and [4] is intelligible because I may have only found out about Nouméa while doing research on French colonies the previous week. In neither case does MP arise or is there any conceptual confusion about the nature of belief.

MP supports TM in my view because it illustrates the conceptual connection between occurrent judgement and belief. It shows how judgement underpins belief.[10] By this I mean that occurrently judging that *P* conceptually entails believing that *P*. A person who judges that Nouméa is the capital city of New Caledonia accepts a particular claim about the world to be true, and in doing so believes. I am not suggesting that such a belief would be the causal result of the judgement that *P*: that judging that *P* causes the belief that *P* to arise. In my view, the relationship is comparable with the one between losing one's temper and shouting in a loud aggressive way, and yelling. I would not say that losing one's temper and shouting *causes* the yelling, but that it *conceptually entails* it—that is what it means to yell.

This analogy, while hopefully illustrative, is not perfect however because beliefs can sometimes arise in isolation to conscious judgement. I may believe that *P* unconsciously, for example. My behaviour may suggest to others that I believe that *P*—and let us assume that I do believe that *P*—even though I never judged that *P*. This is, however, quite compatible with my contention. All I am claiming is that judgement conceptually entails believing, not that believing conceptually entails consciously judging.

It is for this reason that it would be so strange for someone to occurrently judge that Nouméa is the capital city of New Caledonia and then claim that they do not believe that it is.[11] Such a person would appear to be conceptually confused about what it

[10] Following Cassam (2010), I understand judgements to be conscious *mental actions*, which are related to beliefs, which are typically more stable, longstanding, *mental states*. Cassam would disagree, however, that judgement and belief go together with the kind of congruity I am proposing here. Such a discussion is beyond the scope of this paper, however, which is after all not a thorough assessment of whether TM is a successful method for self-knowledge of belief, but instead an attempt to apply it to other propositional attitudes.

[11] It would also be absurd for someone to judge that *P* is the case and claim that they have no beliefs either way about *P*. Consider someone who were to avow the following: 'I judge that Nouméa is the capital city of New Caledonia, but I neither believe it or disbelieve it'.

means to believe something. It is this conceptual relation that I think justifies the claim that occurrent judgement can be a guide to what one currently believes—the question 'Do I believe that *P*?' is transparent to the question 'Is *P* true?'[12]

My proposal in the present paper is that belief provides a model for extending TM to other attitudes. What I will argue is that in order to apply TM to other mental states, we must look to the concepts that underpin those mental states. Judgement is a guide to what one believes because judging that *P* conceptually entails believing that *P*. Extending TM to desires, intentions, wishes and so on requires that we attend to the concepts that underpin those mental states. Doing so, I argue, reveals a series of outward-directed questions that can be attended to, in order to know what one desires, intends, wishes and so on. Call this the conceptual approach to extending TM. In what follows, I support the conceptual approach by showing how it generates Moore-Paradoxical sentences that are analogous to the case of belief.

In doing so, I will remain as neutral as I can with respect to a debate in the literature among TM theorists about the role that inference versus the role that rationality plays with respect to these outward-directed questions. Moran (2001, 2012) and Byrne (2005, 2018) for instance both think that one can know whether one believes that *P* by determining whether *P* is the case, yet they disagree about how the process works. Moran, for example, thinks that in order to determine whether *P* is the case, engagement with our rational faculties is required; whereas Byrne rejects Moran's appeal to rationality and claims that one is justified in self-attributing a belief that *P* by making an inference from the response one gives to a worldly question about *P*. He says, 'one (allegedly) comes to know that one believes that *p* by inference from the premise that *p*' (2018, p. 15). What Byrne has sought to show, in his recent work, is how an inference from a worldly premise to a conclusion about a psychological question is warranted.[13]

Despite these differences, Moran and Byrne are both TM theorists: they both agree that answering an outward-directed question such as 'Is *P* true?' can yield self-knowledge of what one believes. While my own sympathies lie with Moran's rationalistic approach here, I will not attempt to weigh into this debate in what follows. I will limit the focus of the paper to the task of determining what the outward-directed questions for non-belief mental states would look like. How we should think about rationality and inference with respect to these outward-directed questions is an important issue to address, but one that must await another occasion. First, let us look at some reasons for thinking that such questions cannot be found.

3 The Generality Problem for TM

Those who think that there is a Generality Problem for TM do not claim that following TM will *never* yield self-knowledge. Carruthers (2011, pp. 83–84), for example, is critical of TM and yet thinks that following it will *sometimes* yield self-knowledge, such as in the case of one's perceptual beliefs. He thinks that someone can achieve self-knowledge of their belief that a toy is broken by *judging* that they are *seeing* a broken

[12] See Schwitzgebel (2010) for a discussion about how to classify cases where there is an apparent mismatch between what a person believes and what they occurrently judge to be the case.

[13] For a recent critique of Byrne's approach, see Boyle (2019, pp. 1019–1024).

toy. (This would count as adhering to TM because one is attending to the content of the belief, which is the broken toy.) Carruthers does not think that much follows from this admission, however. He claims that '[it] is unclear how such accounts [TM] could generalize to many other types of attitude besides belief and judgment' (2011, p. 84).[14]

Shaun Nichols and Stephen Stich, like Carruthers, think that TM cannot be generalised to account for the self-knowledge one can have of one's non-doxastic mental states (e.g. one's desires and intentions). With respect to questions such as 'Do I desire that *P*?' and 'Do I hope that *P*?', they argue that, '[t]here is no plausible way of recasting these questions so that they are questions about the world rather than about one's mental state' (2003, p. 194). David Finkelstein makes the same criticism in the following passage:

> it is difficult to claim that the self-ascription of belief provides a model of self-knowledge that can be used in order to understand our awareness of our own, say, desires because there seems to be no "outward-directed" question that bears the kind of relation to "Do I want X?" that the question "Is it the case that *p*?" bears to "Do I believe that *p*"? (2003, p. 161).

The common thread running through these criticisms is the idea that TM is only applicable to belief because there do not exist any outward-directed questions that one can attend to in order to know what one desires, intends, wishes and so on. Since these authors do not offer any reason in principle why this is the case, a response to the Generality Problem can gained by providing examples of such outward-directed questions.

While there have been some significant attempts to address the Generality Problem in the last few years, it is by no means fully resolved. Akeel Bilgrami, for example, in his book *Self-Knowledge and Resentment*, largely focuses on belief and desire, and says that '[e]xtending the account given here to the other intentional states and eventually to qualitative states of mind is an important task but must await another occasion' (2006, p. 1). And Richard Moran—another one of TM's leading proponents—holds that, while belief is 'representative' (2012, p. 214) of the way in which we acquire knowledge of our attitudes in general, he (much like Bilgrami) does not provide a detailed account of how such an extension may be realized.

In the past few years, there have been several attempts to fill this lacuna. For example, Fernández (2007, 2013) provides an account of how TM can be extended to desire. Paul (2012) and Boyle (2011) provide an account of how TM can be extended to intentions. Barz (2015) does the same for wishes. And Alex Byrne (2018) has recently accounted for several different kinds of mental states, including sensations, desires, intentions and emotions. In what follows, I argue that while these accounts make important inroads into solving the Generality Problem, they are not completely successful. I advance my own view, the conceptual approach, which I argue

[14] It is interesting to compare Carruthers' criticism of TM: that it is *only* applicable to sensory-based beliefs, with Dorit Bar-On's criticism of TM: that there is *no way* to extend TM to sensory-based beliefs (see 2004, p. 118). These distinct criticisms illustrate how differently TM has been represented in the literature.

provides a more adequate solution to the Generality Problem. I start with the attitude of desire, before looking at intention, and then finally wishes.

4 The Analogous Question for Desire

I claim that one can acquire knowledge of what one desires by attending to an outward-directed question—analogous to the way that one can know whether one believes that *P* by attending to the question 'Is *P* true?'[15] I begin by offering a critical discussion of some recent attempts to extend TM to desire, before then advancing the conceptual approach.

4.1 Some Recent Approaches to Desire

Fernández (2013) argues that self-knowledge of our desires can be accounted for by appealing to TM. He claims that we normally 'form beliefs about our desires on the basis of our grounds of those desires' (2013, p. 86). This means that we do not answer questions such as 'Do you desire to ϕ?' or 'Do you want it to be that case that *P*?' by searching our minds for the presence of a desire.[16] Instead, he argues, we *look outward*, to the 'intentional object of the desire' (2013, p. 87). Let us consider Fernández's own example to illustrate how this approach works. Suppose I am asked, 'Do you desire to go to the party?' On Fernández's account, I can know that I do have this desire by considering 'whether going to that party might be fun' (2013, p. 87). So, if I judge that the party will be fun, then I should believe that I do desire to go to the party. In the same way, Fernández claims that I can know that I desire to get a drink by considering 'whether I feel like having one' (2013, p. 87).

Although I agree with Fernández that we do acquire self-knowledge of our desires by following TM, I do not believe that he succeeds in his attempt to extend TM to desire. This is because he does not provide us with a model, or formula, for extending TM to desire. If the outward-directed question for belief is 'Is *P* true?', we may ask of Fernández's account: 'What is the analogous question for desire?' It cannot be, after all, that the question 'Is ϕ-ing fun?' (to take Fernández's own example) can be used as a general model for acquiring knowledge of one's desires—as many desires will not take this form. I may desire to give $100 to disaster relief, for example, but I would not say that I can know that I have this desire by attending to the question 'Is giving to disaster relief fun?' Neither do I think it is the case that the question 'Do I feel like ϕ-ing?'—to draw upon Fernández's other example involving a drink—gives us the right answer. This is because the question 'Do I feel like ϕ-ing?' features the word 'feel', which in this context is describing 'desire' by another name. If an outward-directed question for desire is to be sought, it should not involve a desire-like mental state. It is true that doing so would produce Moore-Paradoxical sentences, as in the case of 'I feel like getting a drink, but I do not desire to'. However, this is unsurprising given that these

[15] I will only be concerned with how we know our occurrent conscious desires. As Krista Lawlor points out, we sometimes become aware of our unconscious desires by 'inference' (2009, p. 49), e.g. by testimony or therapy. This method of achieving self-knowledge of our desires would not conform to TM. This shows that TM is not the only way one can know what one desires.

[16] In what follows, I consider the terms 'desire' and 'want' to be analogous.

two conjuncts are essentially identical. This is unlike when TM is applied to belief, where we said that judgement can be a guide to belief. In that case, judgement and belief are not the same thing. The former underpins, or conceptually entails, the latter. So, while Fernández is right to focus on the transparency of desire, his outward-direction question does not capture the concept of desire. By this I mean that the question Fernández provides does not inform us about the nature of desire, in the same way that judgement informs us about the nature of belief. Thus, I do not think that he has provided us with an adequate explanation of how TM is applicable to desire.

Byrne (2012, 2018) proposes an alternative explanation for how TM can be extended to desire. He does so by advancing the epistemic rule DES, which says: 'If ϕing is a desirable option, believe that you want to ϕ' (2018, p. 161). Put another way, DES says that one can know whether one desires to ϕ, by attending to the outward-directed question 'Is ϕ-ing a desirable option?' According to DES, then, I can know whether I desire to go windsurfing by determining whether windsurfing is a desirable option. Byrne thinks that knowledge of 'one's desires is typically obtained by trying to follow DES' (2012, p. 177).

Byrne concedes that DES is not an infallible guide to what one desires. He grants that cases such as the following are possible. Suppose that one of Tim's favourite things to do on a Sunday morning is to take his dog Iggy for a walk through his neighbourhood. Now, does it follow that if on Sunday morning Tim judges that taking Iggy for a walk is a desirable option that he should believe that he wants to take Iggy for a walk—as is suggested by DES? Byrne suggests not. Tim might have had one too many glasses of wine the night before, and the idea of getting out of bed and going for a walk makes his head spin. Despite the possibility of such cases, Byrne still thinks DES is '*practically* self-verifying' (2012, p. 178). He grants that this does not necessarily make DES a good rule, but he does think that the 'burden of proof should be on those who think it is not' (2012, p. 178).

To my mind, DES faces several problems that go much deeper than the rule's tendency to yield false beliefs about what one desires. First, I think that the way in which Byrne connects 'finding something a desirable option' and 'wanting' to perform that action is problematic. To see why, let us note that Byrne explicitly states that he understands the terms 'want' and 'desire' as equivalent (2018, p. 158 ft. 5). This allows him to connect (1) the believing that one wants to ϕ, with (2) the knowledge that one desires to ϕ. But he also thinks that 'desirable option' cannot be identical with what one desires, as by Byrne's own lights one can find an option desirable and yet not desire it. This leaves Byrne with a difficult question: 'What does it mean to find an option a desirable one, if not simply the fact that one desires to do it?'

Byrne acknowledges that 'desired option' cannot simply mean that one desires an option, as that would make DES circular (see 2018, p. 162). It would also make DES superfluous—if one had to know what one desires, to follow DES, then it would not make it a very useful rule. On the other hand, if we deny that 'desirable option' simply means 'to desire something', then Byrne is left with the problem of specifying what it means to find something a desirable option. What is it for Tim to find the walking of his dog a desirable option, for example, if not the fact that he desires to walk his dog? Byrne may be right that this can be accounted for without appealing to desire, but his locution 'desirable option' is difficult to understand in isolation from the act of desiring.

To make my point more explicit, let us imagine that we create a rule for belief which applies a similar structure to DES. Suppose I claim that one can determine whether one believes that *P* by attending to the question 'Is *P* believable?' For example, suppose I claimed that I could know whether I believe that Jane is honest by attending to the question 'Is the proposition "Jane is honest" believable?' As a way of finding out what I believe, the rule would not be false. However, it would not be a very useful rule, as determining whether *P* is believable is just another way of asking whether I believe that *P*. I think that the same flaw is present in DES. It is not that DES fails to produce MP sentences—it does. And it is not that DES will fail to produce self-knowledge if followed—it may. It is that DES only does so because 'desirable option' is just another way of expressing one's desire.[17]

DES does not get to the concept of desire, in the same way that judgement gets to the concept of belief. Thus, I do not think DES succeeds, as an application of TM to desire.

4.2 The Conceptual Approach to Desire

The central principle that guides my approach to extending TM to desire is the idea that we must look to the *concept* of desire. What does it mean to desire something, then? Although there is controversy, with respect to the question of what desire is, most philosophers would agree that desire motivates action (see, e.g., Smith 1994; Schroeder 2004; Sinhababu 2017a). Michael Smith for example says, 'a desire that p tends to endure, disposing the subject in that state to bring it about that p' (1994, p. 115). This means that if S has a desire for a cold beer on a hot day, then she will typically be motivated to bring about the state of affairs where she is drinking a cold beer—say by walking to the fridge. In cases where one desires something that one does not have the power to bring about, such as when one has a desire for their sports team to win, one will be disposed to *prefer* their sports team to win. It is, of course, not always easy or possible to act on one's desires, such as when one has a desire to ask a question in a crowded lecture hall, but is overcome with nervousness; or when one desires to see all three movies that are playing at a cinema at the same time, but can only see one (see footnote 19).

Since desire motivates behaviour, it is a type of mental state that typically instils a pleasurable experience in a subject who imagines a certain state of affairs occurring. As Neil Sinhababu points out, desires are the sorts of mental states that 'cause pleasure when we sense or imagine them being satisfied, and displeasure when we sense or imagine them unsatisfied' (2017b, p. 95). If I am thirsty, then the idea of getting a drink

[17] It has been pointed out to me that some people use the term 'believable' differently to the way that I do here. Some use the term to refer to the upper threshold of epistemic possibility. For example, one might say 'It's not believable that the LA Kings will win the Stanley Cup' to mean that it is very unlikely the LA Kings will win. Given this interpretation, the sentence 'It's not believable that the LA Kings will win the Stanley Cup, but I believe they will' does not sound very Moore-Paradoxical. This is because there is a significant gap between finding something believable, in this epistemic threshold sense, and actually believing it. This would, therefore, not make it a good guide for determining what one believed. If we were to interpret DES in this same way, such that 'desirable' referred to an objective threshold of what is typically desired by people, then DES would be rendered even more implausible than I have suggested. This is because the connection between what one desires and what one deems desirable (in this objective threshold sense) is not a very congruous one. For example, a smoker may deem it 'undesirable' in this sense to have a cigarette, yet still have a desire for one.

causes *pleasure* when I imagine this state of affairs eventuating; similarly, if I am told by my boss that I will have to work on the weekend, I will feel *displeasure* when I think about how I will have to cancel the hiking day trip I have been planning for weeks. On other occasions, desires are more closely aligned with satisfaction or relief. For example, when my desire for a loud car alarm to stop going off is fulfilled, I am likely to feel satisfied or relieved, rather than overly joyous.

All of this points to the concept of a desire: when someone consciously judges that ϕ-ing (or the occurrence of P) will bring one pleasure or satisfaction, then I think that person has a desire to ϕ (or for P to occur). By this I mean that occurrently judging that ϕ-ing (or the occurrence of P) will bring one pleasure or satisfaction *conceptually entails* desiring to ϕ (or that P occurs). Importantly, I am not saying that the converse is true: if someone desires something, then they have judged that ϕ-ing (or the occurrence of P) will bring them pleasure or satisfaction. This would rule out unconscious desires, which are possible.

By reflecting upon these properties of desire, I think that an outward-directed question for desire emerges. My proposal is that one can know what one desires by attending to the question: 'Would ϕ-ing (or the occurrence of P) bring me pleasure or satisfaction?' Self-knowledge, on my account, is not achieved by *looking inside* and noticing the presence of a certain desire, but by judging whether a certain state of affairs will bring one pleasure or satisfaction. We can formulate this idea as follows.

> **The Conceptual Approach to Desire:** The question 'Do I desire to ϕ?' (or 'Do I desire that *P?*') is transparent to the question 'Would ϕ-ing bring me pleasure or satisfaction?' (or 'Would P's occurrence bring me pleasure or satisfaction?').[18]

An example will help to explicate the view. Suppose James is asked, 'Do you desire to go to the karaoke bar tonight?' According to my account, James can gain knowledge of whether he does by attending to the question, 'Would going to the karaoke bar bring me pleasure or satisfaction?' If he judges that it would, then he should attribute such a desire to himself. Does it follow that James will attempt to fulfil this desire? No, he may need to study for an exam the next day (that he desires to pass), and so may express his disappointment to his friends who are going: 'sounds like it would a great night, but sadly I have other commitments'. It may be objected that this shows James does not actually desire to go, but I would say that attributing such a desire to James explains why he is disappointed that he cannot go. Call this the *conceptual approach* to extending TM to desire.

An advantage of the conceptual approach is that it is applicable to the different varieties of desires that are commonly distinguished between in the literature such as instrumental and non-instrumental desires (see Fernández 2013, pp. 83–84). Instrumental desires are desires we have for states of affairs which we do not desire for their own sake, but rather because we believe they will lead to something we do desire. Consider, for example, James's desire to take the bus to the karaoke bar. Suppose he only has this desire because he desires to go the karaoke bar and believes that catching

[18] I use these two forms to distinguish between desires which bring about actions ('ϕ'); versus those that are about states of affairs ('P').

the bus will get him there. His desire to catch the bus is instrumental because he does not desire it in itself—we may suppose that James believes that buses stop frequently, are uncomfortable and unreliable. On the conceptual approach, James could know whether he has this desire by determining whether catching the bus would bring him pleasure or satisfaction. If he judges that it would, which he seems to have good reasons to conclude, then he should believe that he desires to catch the bus. This will be the case even though the bus ride will not bring him much pleasure or satisfaction. He only catches the bus because he believes it will help him fulfil his desire.

Non-instrumental desires, on the other hand, are states of affairs we desire in themselves. An urge, such as a desire to eat a freshly baked bagel, would be an example of a non-instrumental desire. This is because one would typically desire it for its own sake. According to the conceptual approach, to know whether one desires to eat the bagel, one should determine whether doing so will bring one pleasure or satisfaction. The conceptual approach is also applicable to urges that one may be ashamed of, or preferred one did not have. If one judges that smoking a cigarette will bring one pleasure or satisfaction, for example, then one should self-ascribe the desire to smoke a cigarette. This will be the case even if one is ashamed of, or wishes they never had, this desire.

In contrast to these 'basic' non-instrumental desires, there also exists 'non-basic' non-instrumental desires—following Fernandez's (2013, p. 83) terminology. Non-basic desires are not the kind of desires that spontaneously (and passively) arise—they are more complex in their nature. Consider, for example, Axel's desire to be a professional musician. We may imagine that Axel has this desire not because it will bring him fame or fortune, but for its own sake. This desire is unlike his desire to have a drink on a hot day because it is 'judgement-sensitive' as Scanlon (1998, p. 21) puts it. This means it is a desire that is subject to rational evaluation. The conceptual approach explains how Axel could gain self-knowledge of this desire. If he judges that becoming a professional musician will bring him pleasure or satisfaction, he should believe he desires to become one.

What argumentative support can be provided for the claim that TM can be applied to desire in the way I that have described? As with the case of belief, I think that we can appeal to Moore's Paradox. For example, it would be absurd for someone to utter, or imagine, the following sentence: 'Going to the karaoke bar would bring me pleasure or satisfaction, but I do not have a desire to go there'.[19] Now, as we found with the MP sentences for belief, this avowal need not be necessarily false. This is because it involves a judgement about what will bring one pleasure or satisfaction, and a self-attribution of desire—the former occurring on the first-order level and the latter on the second-order level. The sentence does sound absurd, though. Anyone uttering it would

[19] It may be objected that we often imagine things will bring us pleasure or satisfaction, and yet do not *want* to do them. For example, there may be three movies showing simultaneous at a cinema complex: *Jaws*, *Alien* and *Scarface*—all movies I think I would enjoy. Suppose I can only see one and come to pick *Jaws.* I would not say after making this choice that I lack a desire to see *Alien* or *Scarface*. I would say that I still believe that both would bring me pleasure or satisfaction, and so I think I have a desire to see both. However, since I believe that *Jaws* would be more pleasurable (or enjoyable) than the two other movies I decide to see it—and may even say that I do not want to see *Alien* or *Scarface*. It is important to note that the sense of 'want' in this context should be understood as shorthand for preference and should not be construed as denoting a lack of a desire to see *Alien* or *Scarface.*

seem to be unsure about what it means to desire something. I concede that it would not be absurd to say, 'Going to the karaoke bar would bring me pleasure, but I do not *want* to go' if the phrase 'want to' denoted the speaker's intention not to go. But I think that if we really think about what the speaker is saying, it would be strange for her to judge that going would be pleasurable, but then claim that she did not have some desire to go. If she did not have such a desire, then why would she judge that going would be a pleasurable experience? So, I think that the conceptual approach does generate MP sentences.

In contrast to the other accounts we have looked at, however, the conceptual approach does not simply feature a desire by some other name—such as one's feeling that *P* or one finding *P* a desirable option. The conceptual approach, rather, appeals to what it is that underpins (and conceptually entails) desire—namely pleasure and satisfaction.

In advancing the conceptual approach as a guide to self-knowledge of desire, I do not mean to suggest that it provides an infallible guide. The conceptual approach is compatible with self-deception, as well as other forms of misclassification of desire. The following is possible on the conceptual approach. Imagine that Sam's friend asks him whether he desires to go on a date with Penny. We can imagine that Sam follows the conceptual approach by judging whether going on a date with Penny will bring him pleasure or satisfaction. Let us suppose that Sam judges that it will not, and thus does not attribute the desire to go on a date with Penny. Now, does it follow that he does not desire to go on a date with Penny? Clearly not. It may be that Sam is self-deceived. He may have an unconscious desire to go out with Penny. However, he may find it too difficult to admit that doing so will bring him pleasure or satisfaction, because of his lack of confidence or social anxiety.

Despite such exceptions, I still think the conceptual approach succeeds. No account of self-knowledge, after all, should possess the property of infallibility. The outward-direction question 'Would ϕ-ing bring me pleasure or satisfaction?' can be used as a guide to what one desires because it generates MP sentences analogous to the case of belief. It does so, moreover, without simply involving a desire state by a different name—it invokes the concept of desire.

It may be objected that my account still goes awry because it is easy to imagine cases where one judges that ϕ-ing will bring one pleasure or satisfaction, where a desire to ϕ is not present. Consider Robert Nozick's famous experience machine thought experiment, which involves a machine that if plugged into 'would give you any experience you desired' (1974, p. 42). Hooked up to the machine one may, for example, experience reaching the top of one's chosen field, meet the love of one's life or be extremely wealthy, while in reality one would be floating in a tank—even though one would not be aware of this while in the tank. While the prospect of plugging in may initially seem enticing, there is a big catch: the machine 'limits us to a man-made reality' (Nozick 1974, p. 43).

Even though plugging in would mean that one could experience large amounts of pleasure and satisfaction, most people who come across Nozick's though experiment claim that they would not plug in. Most of us would prefer to live our lives in the real world with all its imperfections. However, this produces a problem for my view in the following way. Imagine that I ask, 'Will plugging into the experience machine bring me pleasure or satisfaction?' It seems clear that plugging in will—after all, Nozick stipulates we can experience any desire of our choosing. So, it seems that since I would have all my desires fulfilled, I should reply 'yes'. However, this raises a problem, for it

seems that my answer to the 'mind-directed' question 'Do I desire to ϕ?' is 'no'. I do not desire to be plugged into the experience machine. So, the experience machine case seems to bring into question the conceptual connection between desire on the one hand, and pleasure and satisfaction on the other.

To respond to this objection, it is important to first acknowledge that we are focused on the question of desire, not on what one would choose to do. There are, after all, times where fulfilling our desires would lead to consequences we do not desire, so we do not choose to do them. Partying all night at the karaoke bar may be desirable for an undergraduate student, but it may prevent her from fulfilling her desire to pass her exam the next day. A strong-willed student is the one who overcomes such a desire and gets a good amount of rest. Analogously with the experience machine case, I do not think that we should deny that the person, who judges that going into the machine will bring them pleasure, has a desire to plug in. What I think we should say is that such a person would not *choose* to go in, even though they believe that doing so would lead to pleasure, and thus the fulfilment of their desires.

The reason that I would not choose to go in, for example, is that I have a desire for truth, a desire to be there for my friends, colleagues, students or anyone else that may need my help. Helping a friend through a difficult period of their life is not achieved in the experience machine because such a friend would have never experienced a difficult period—indeed they would have no experiences at all (they would not even exist). So, helping them would be of no real benefit. In the machine, I would believe that I helped them, and had made a real difference to their life, but this belief would be false. The fact that all of my other desires of this sort would also fail to be really fulfilled would explain why I would choose not to go in, even though I would experience pleasure in the machine.

So, I do not think that the experience machine case undermines the conceptual approach to desire. Unlike with belief, it is not uncommon for us to have conflicting desires. Selecting what the best course of action is when one has conflicting desires can be hard of course. But this is not an issue for the conceptual approach to desire.

5 The Analogous Question for Intention

I will now show how the question, 'Do you intend to ϕ?' can be recast as an outward-directed question, in the same way that the question 'Do you believe that *P*?' can be recast as the question 'Is *P* true?' First, I offer a critical discussion of two recent attempts to extend TM to intention, before presenting the conceptual approach.

5.1 Some Recent Approaches

Byrne (2018) argues that TM can be extended to intention by appealing to his epistemic rule INT. Byrne defines this rule as follows: 'If you will ϕ, believe you intend to ϕ' (2018, p.169). Another way of putting this is to say that one can know whether one intends to ϕ by attending to the question 'Will I ϕ?' Let us consider an example to illustrate how this approach works. Imagine that Robert intends to go to the Casino on Saturday night. According to INT, Robert can acquire knowledge of his intention by attending to the

question 'Will I go to the casino on Saturday night?' If Robert judges that he will, then he should believe that he intends to go the casino on Saturday night.

INT seems plausible at first pass since one often intends to do what one judges/thinks/believes that one will do. It seems that if I judge that I will go to the party, then I should self-ascribe the intention to go. On closer inspection, however, the rule seems flawed. This is because there are times when one does *not* intend to do what one believes one will do. Consider the following example. Suppose Jane is a graduate student preparing to give her first talk at a large conference. Although Jane is excited to give the talk, she is also aware of how much public speaking terrifies her. She believes that she will be nervous, that her voice will tremble, and that she will get overwhelmed. She does not want this of course and is actively trying to do all she can to avoid being nervous—she is reading books on public speaking and practising in front of the mirror. Nevertheless, she believes that she will be nervous. If she follows INT, she should believe that she intends to be nervous. This result seems false, given how actively Jane is trying to avoid this happening.

Consider another example—one that Byrne himself discusses—that is raised by Elizabeth Anscombe. Anscombe imagines a case where a poorly prepared student is about to take an examination and thinks to himself 'I am going to fail in this exam' ([1957] 2000, p. 2). In such a case, we can think of the student's thought as a prediction of what he will do, rather than an intention to fail the exam. We can imagine that in the moments leading up to the exam the student is doing all he can to remember the content that he is being tested on. In these final moments before the exam, he has the intention to pass, but being realistic, he believes he will fail. Following INT will not give the right answer here either: if the student believes he will fail the exam, INT says that the student should self-attribute the intention to fail the exam. This seems false.

Despite the existence of such cases, Byrne still maintains that INT is a good rule because it is '*practically* strongly self-verifying' (2011, p. 219). Byrne explains the meaning of this locution by saying: 'if one reasons in accord with the schema (and is mindful of defeating conditions, for instance the one just noted), then one will arrive at a true belief about one's intention' (2011, p. 219). While I agree with Byrne that INT will *sometimes* yield true beliefs about one's own intentions, this does not make it a good rule in my view. As the cases we have considered show, INT does not seem to be able to account for the fact that someone who intends to ϕ will not only *believe* that they *will* ϕ, they will also be prepared to act in various ways to bring about ϕ. INT requires one to take a predictive stance towards one's course of behaviour—meaning that one must focus on whether a certain state of affairs *will* happen to oneself. This means that MP sentences are not produced in cases such as the following: 'I will get nervous during my talk, but I do not intend to'—a sentence which is quite natural. Given that INT does not produce MP sentences in such cases, I do not think it succeeds as a way of applying TM to intention.

An alternative approach to Byrne's—one that gets closer to the concept of intention—is given by Sarah Paul. She argues that 'we can come to know what we intend by making a decision about what to do and self-ascribing the content of that decision as our intended action' (2012, p. 327). On Paul's account, then, the question that is transparent to 'Do you intend to ϕ?' is: 'Have I decided to ϕ?' The addition of this deliberative element seems to capture something important about what it means to have an intention—something that is absent from Byrne's account.

If we recall the case involving Jane, Paul's account gives us the right answer. Since Jane has not decided to get nervous when she is giving the talk, she should not self-attribute the intention to do so. Similarly, with the student who is about to fail the exam. On Paul's account, he should not attribute the intention to fail the exam since he has not decided to. Given that Paul's account gets closer to the concept of intention, it also produces 'absurd' sounding MP sentences. If someone were to say, 'I have decided to go to the party, but I do not intend to', they would be met with the same confusion as someone who said, 'Nouméa is the capital of New Caledonia, but I do not believe it is'.

Even though Paul's account gets closer to the concept of intention than Byrne's, I do not think that it fully succeeds. This is because I think that deciding to ϕ will only be able to account for a subset of our self-ascriptions of intention. Paul claims that 'deciding to ϕ is normally sufficient to count as intending to ϕ' (2012, p. 343). In my view, however, deciding to ϕ will only be a way to gain knowledge of one's intentions in cases where one forms the intention for the first time—that is, I do not think that Paul's approach captures the process of coming to know one's already formed intentions. For example, if I am asked whether I intend to go overseas during my summer vacation, and I have already booked my trip, then it does not seem like I need to decide whether I will go, in order to know whether I intend to go. This is because I have already decided that I will go—my mind is already made up.

5.2 The Conceptual Approach to Intention

To extend the conceptual approach to intention, we must first ask: 'What does it mean to intend to do something?' Although there are controversies surrounding the answer to this question, most philosophers would say that to intend to do something is to actively attempt to bring about a certain state of affairs. As Michael Bratman says, 'an intention to act is a complex form of commitment to action, a commitment revealed in reasoning as well as in action' ([1987] 1999, p. 110). An intention, thus, is not just an event in the future that one believes will happen to oneself. As Amir Saemi states, an 'agent needs to have some commitment to execute his/her intention' (2015 p. 202). My intention to fly to Hawaii for the summer is not just a belief about what *will* happen to me in the summer; but rather, it is a series of commitments of mine that relate to my goal of bringing about this state of affairs—namely, flying to Hawaii. If I am really committed to going to Hawaii for the summer, then I should make genuine attempts, for example, to book flights, search for accommodation, and put money aside. I need not succeed in accomplishing these tasks, of course, but I should make legitimate efforts to do so, if I am to be said to truly intend to go to Hawaii for the summer. Someone who merely has the thought that going to Hawaii in the summer would be fun, while making no legitimate attempts to bring about the state of affairs where they are in Hawaii, would not be committed to going in the sense I have in mind here. As a result, I would say that they do not possess an intention to go to Hawaii—even though they may have a desire to go.

An intention in this sense is not just a promise to ϕ or a feeling of obligation to ϕ. One may have a feeling, compulsion, obligation or have promised to ϕ but make no real attempt to act. Charles may feel obligated to—or feel like he ought to—give money to the disaster appeal to support victims of a bush fire but make no serious attempt to do so. The person who is committed to ϕ, on the contrary, makes a legitimate attempt to ϕ, and believes that his attempt will lead to, or have a chance of leading to the occurrence of P. Reasoning and planning here is important.

By reflecting upon these properties of intention, I think that an outward-directed question for intention arises:

The Conceptual Approach to Intention: The question 'Do I intend to ϕ?' is transparent to the question 'Am I committed to ϕ-ing?'

Self-knowledge, then, on my view, in not achieved by looking inside and noticing the presence of a certain intention. Instead, it is achieved by attending to the intentional object of one's intention and then determining whether one is committed to bringing that state of affairs about. This may involve making a decision about ϕ-ing, but it may also require one to determine: whether one still wants to ϕ after previously having decided to ϕ, whether ϕ-ing is possible, or whether ϕ-ing is the best thing to do. Determining whether one is committed to ϕ-ing also features a normative component. A rational actor, after all, ought not to be committed to an action they believe to be impossible to perform.

In claiming that there is an important connection between commitment and intention, I am following several other authors, such as Matthew Boyle who says: 'My intention is a kind of commitment to ϕ,' (2011, p. 234).[20] The way that I would characterize this relationship, however, is not to say that the two are identical, however; rather I would say that commitment *underpins* intention—in the same way that judgement underpins belief. Being committed to ϕ-ing conceptually entails intending to ϕ.

Support for this claim, like in the case of belief, can be gained by appealing to Moore's Paradox. Consider the sentences 'I am committed to passing the exam, but I don't intend to'; and 'I am committed to going to the party, but I don't intend to'. These sentences sound very Moore-Paradoxical. The reason for this, in my view, is because commitment underpins (or conceptually entails) intention. The conceptual approach produces MP sentences, therefore, because it gets the concept of intention right.[21] Compare this with Byrne's approach to intention. On his account, one can determine what one intends to do by determining what one will do. It is important to note the MP sentences are not generated from his view, for example: 'I will fail the exam, but I don't intend to'; and 'I will lose my job after my latest indiscretion, but I don't intend to'. There is nothing absurd about such sentences. This suggests that the connection between what will happen to one and what one intends to do is only contingent. Given that the conceptual approach does produce MP sentences, as was shown, I think that it succeeds in accounting for the self-knowledge we can have of our intentions.

It may be objected that there are times where MP sentences fail to arise with respect to the conceptual approach, so let us address that concern. Imagine that John promises, out of politeness, to attend Jamie's Christmas Party after he invites him. Suppose further that the very moment after this occurs, John instantly feels dread. The last thing he wants to do is be at a party that is hosted by Jamie—a person who he secretly cannot stand. Still, he sighs and reflects: 'A promise is a promise'. Now, might John think:

[20] See also Bilgrami (2006), who discusses the connection between intention and commitment.

[21] As in the case of belief and desire, the conceptual approach should not be understood as an infallible guide to intention.

'Well, I am committed to going to the party now, but I do not intend to go'? Such a thought seems to indicate that a conceptual connection is not present between intention and commitment, as the sentence does not seem paradoxical. Thus, my strategy would seem to have a fatal flaw.

I would respond to this objection by rejecting the claim that John lacks an intention here. If we are talking about the self-ascription of an intention, then I think it is hard to see how he would not have one. After all, he is planning to go to the party, he has decided to go, he believes he will go, he may have found a babysitter to look after his children and so on. This certainly sounds like someone who is intending to go to the party. Now, it is true that John might not look forward to going, or regret accepting the invitation, but I would still say he intends to go. If we do not attribute such an intention, how would we explain his attendance at the party? After all, we often intend to do things we believe we will not enjoy. Thus, I think that once properly understood, John's sentence actually is Moore-Paradoxical, and the objection can be answered.

6 The Analogous Question for Wishes

Finally, I will consider how TM can be extended to wishes. I will begin by looking at a solution that Barz (2015) has recently proposed. While Barz claims that there is no outward-directed question—'no question about external matters' (2015, p. 2016)—that one can attend to in order to know what one wishes, he does think that a 'Byrne-like' rule can be provided, in order to account for the self-knowledge one can have of one's wishes. He claims that 'If the lacuna that is determined by 'If only *p* were the case!' exists, believe that you wish that *p*' (2015, p. 2017). Barz refers to such a sentence as a 'postulation' (2015, p. 2017) and claims that one can know whether one wishes that *P*, by determining whether one *endorses* such a postulation. (Barz claims one would need to endorse the postulation because unlike propositions, postulations are not taken as true or false.)

Barz thinks that his approach can be supported by the fact that it generates MP sentences. He thinks it would be absurd, for example, for someone to think or say 'If only Congress would pass more restrictive gun laws! But, well, I do not wish that Congress would pass more restrictive gun laws' (2015, p. 2017). In my view, while it may be true that Barz's account generates MP sentences, it does not follow from this result that his account succeeds. This is because the sentence that he claims can be used as a guide to what one wishes, 'If only *P* were the case!', takes the form of the past English subjunctive—a form used in sentences that describe false or unlikely states of affairs. The reason that this is problematic is because when one employs the English subjunctive in the form that Barz does, one would already be aware that one has the wish. Thus, to use this sentence to know what one wishes is circular. This does not render his rule false, but it does not make it a very useful rule.[22]

[22] The English subjunctive can, of course, be used in various other forms. Consider Adam's (Adams 1970, p. 92) famous example, 'if Oswald hadn't shot Kennedy, someone else would have'. This sentence suggests that the speaker believes that Oswald did kill Kennedy and that Kennedy would have been shot by someone else, if Oswald did not do it.

In order to support my contention, let us consider two typical English sentences of the past subjunctive form: 'If only we didn't have this debt' and 'If only we were rich'. When one thinks or avows such sentences, what one is doing is (1) describing a state of affairs that they believe to be false and (2) expressing a wish for the state of affairs to be different. This can be seen by swapping the earlier sentences with versions which feature the word 'wish'. The sentence 'I wish we were rich' is interchangeable with the sentence 'If only we were rich', and the sentence 'I wish we didn't have this debt' is interchangeable with 'If only we didn't have this debt'. Notice that no meaning is gained or lost when we do so—the sentences mean the same thing. What I think this shows is that Barz's sentence for determining what one wishes, 'If only *P* were the case', is just another way to say that one wishes that *P*. It is not surprising, then, that MP sentences are produced on Barz's account: it is because circularity is involved.

However, if the two sentences really do mean the same thing, then should not a response given to one be applicable to the other? Problematically, for my claim, this does not seem to be the case. Imagine the following conversation between two car enthusiasts watching an advertisement for a Tesla Roadster. Imagine that one says, 'I wish I were rich, so I could afford it' and the other person replies by saying 'No you don't, you cannot stand electric cars.' Now imagine that the same person expresses the wish in a different way: 'If only I was rich enough to afford it', and the same response was given: 'No you don't, you cannot stand electric cars'. This response would not seem intelligible. This suggests that the two sentences do not mean the same thing.[23]

I do not think this is a knock-down objection, however. The sentence which features an explicit reference to a wish features the first-person pronoun. In the second sentence, the pronoun does not feature but is rather implied. So, it is not surprising that the response sounds strained. This raises a problem for my view, however. If the sentence involving 'if only' *implies* a wish, then is this not like the relationship between judgement and belief, where I claimed that judgement conceptually entails (and *implies*) belief? And if that is so, might this be a feature, rather than a bug, of Barz's view?

I think that this thought should be resisted. This is because in the case of belief, judgement underpins, or conceptually entails, belief. Barz's proposal does not feature this relationship: The locutions 'I wish' and 'If only' mean the same thing. To be more precise, often one is used over the other for greater emphasis. As the Cambridge Dictionary claims, 'We use *if only* to express a strong wish that things could be different. It means the same as *I wish* but is stronger' (Cambridge Dictionary 2020). This is not the same relationship that judgement and belief have. Judgements are not just stronger beliefs: judgements are mental actions in their own right. Moore's paradox arises because the two are separate. The difference between 'if only I never broke her heart' and 'I wish I never broke her heart' do not suggest different types of mental states or actions; they are rather different ways of expressing the same state, in this case a wish. Thus, I do not think that Barz's strategy works.

Let us now turn to the conceptual approach, which I argue can account for wishes more cogently. Let us first ask, 'What does it mean to wish something?' According to the *Oxford English Dictionary*, to wish something is to 'Feel or express a strong desire or hope for something that cannot or probably will not happen'. Thus, what someone is

[23] I thank an anonymous peer reviewer for suggesting this problem to me.

doing when they claim: 'I wish everyone in the world had access to clean drinking water by the end of 2022' is (1) expressing a desire and (2) judging that the desire is unlikely or impossible to be fulfilled (in this case because millions of people still lack access to clean drinking water).

Given that an account of desire was given in Section 3, all that we need to do in order to account for the self-knowledge one can have of one's wish that *P* is to add the clause that one's desire that *P* is unlikely or impossible. This yields the following account of wishes:

> **The Conceptual Approach to Wishes**: The question 'Do I wish that *P*' is transparent to the question 'Would *P* bring me pleasure or satisfaction and is *P* unlikely or impossible?'

This approach can be supported by noticing how it produces MP sentences. The sentence 'everyone in the world having access to clean drinking water by the end of 2022 would bring me great pleasure, even though it is unlikely to occur; still, I don't wish it' sounds very Moore-Paradoxical. Unlike Barz's account, it does so without simply featuring a wish like state. The conceptual approach, moreover, tells us something about what it means to wish.

7 Conclusion

In the present paper, I have proposed a solution to the Generality Problem for TM. I have shown that one can know whether one has a propositional attitude by attending to an outward-directed question—rather than *looking inside* and detecting the presence of a mental state. I have done so by defending a view I called the conceptual approach. The following table summarizes this view.

It may be objected that these are not really outward-directed questions, as they require one to consult memories, beliefs about what will happen to one, make decisions, and consider future situations. I grant this is true. Yet the same is also true of Evans' original statement about belief. To judge whether there will be a third world war, one must think about the future, consult one's memory about what one has experienced in one's life, and make a judgement. These are internal activities. However, there is a sense in which such questions are outward directed. That is the sense in which we attend to the *intentional object* of the mental state in question to know it. We

Table 1 The conceptual approach to self-knowledge

Inward-directed question	Outward-directed question
Do I believe that *P*?	Is *P* true?
Do I desire that *P*?/ Do I desire to ϕ?	Would ϕ-ing bring me pleasure or satisfaction?/ Would *P*'s occurrence bring me pleasure or satisfaction?
Do I Intend to ϕ?	Am I committed to ϕ-ing?
Do I wish that *P*?	Would *P* bring me pleasure or satisfaction and is *P* unlikely or impossible?

do not look inside and notice the presence, or existence, of the state, like we would notice objects in the world. Our focus is rather on the intentional objects of the states. It is that sense of outward-directed that the aforementioned questions capture.

I provided support for the conceptual approach by showing how the outward-directed questions listed in Table 1 generate Moore-Paradoxical sentences when assented to. This shows that it would be absurd for someone to give a positive response to an outward-directed question and yet fail to self-attribute the corresponding mental state. This supports the thesis that TM can be extended beyond belief.

Although I did not have the space to extend the conceptual approach to all the other propositional attitudes we are capable of holding (e.g., hopes, fears), I have provided a set of criteria for how this could be achieved. First, any outward-directed question that is proposed as a way of knowing what mental state one is in should focus on the concept that underpins that mental state. It should not, for example, involve circularity or involve the mental state by another name. Second, the outward-directed question should yield Moore-Paradoxical sentences. With this general formula in hand, I am confident about the prospect of extending TM to the full range of propositional attitudes. This would, in turn, support the case that TM is a genuine competitor to the inner sense view. The precise details of how such an extension would work, however, like the task of explicating how TM explains first-person authority/privileged access of these mental states, is a topic for another paper.

Acknowledgements Earlier versions of this paper were presented to audiences at the Australasian Association of Philosophy Conference in Adelaide in 2017 and to the Philosophy Society at the University of Western Australia in 2018. I thank everyone who took part in those enlightening discussions. I would like to thank Miri Albahari, Nin Kirkham, Daniel Stoljar, Alex Byrne, André Gallois, Sean Ramsey and Harriet Levenston for feedback and advice on earlier versions of the paper. I also would like to thank an anonymous referee of this journal for the insightful comments.

Compliance with Ethical Standards

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Adams, E. (1970). Subjunctive and indicative conditionals. *Foundations of Language, 6*, 89–94.

Anscombe, G. E. M. ([1957] 2000) Intention. Cambridge: Harvard University Press.

Armstrong, D. M. (1968). *A materialist theory of the mind*. London: Routledge and Kegan Paul.

Ashwell, L. (2013). Deep, dark… or transparent? Knowing our desires. *Philosophical Studies, 165*, 245–256.

Bar-On, D. (2004). *Speaking my mind: expression and self-knowledge*. Oxford: Oxford University Press.

Barz, W. (2015). Transparent introspection of wishes. *Philosophical Studies, 172*, 1993–2023.

Bilgrami, A. (2006). *Self-knowledge and resentment*. Cambridge: Harvard University Press.

Boyle, M. (2011). Transparent self-knowledge. *Aristotelian Society Supplementary, 85*, 223–241.

Boyle, M. (2019). Transparency and reflection. *Canadian Journal of Philosophy, 49*, 1012–1039. https://doi.org/10.1080/00455091.2019.1565621.

Bratman, M. ([1987] 1999) Intention, plans, and practical reason. Stanford: CSLI Publications.

Byrne, A. (2005). Introspection. *Philosophical Topics, 33*, 79–104.

Byrne, A. (2011). Transparency, belief, intention. *Proceedings of the Aristotelian Society, Supplementary, 85*, 201–221.

Byrne, A. (2012). Knowing what I want. In J. Lui & J. Perry (Eds.), *Consciousness and the self: new essays* (pp. 165–183). Cambridge: Cambridge University Press.
Byrne, A. (2018). *Transparency and self-knowledge*. Oxford: Oxford University Press.
Carruthers, P. (2011). *The opacity of mind: an integrative theory of self-knowledge*. Oxford: Oxford University Press.
Cassam, Q. (2010). Judging, believing and thinking. *Philosophical Issues, 20*, 80–95.
Cassam, Q. (2014). *Self-knowledge for humans*. Oxford: Oxford University Press.
Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.
Fernández, J. (2007). Desire and self-knowledge. *The Australasian Journal of Philosophy, 85*, 517–536.
Fernández, J. (2013). *Transparent minds: a study of self-knowledge*. Oxford: Oxford University Press.
Finkelstein, D. (2003). *Expression and the inner*. Cambridge: Harvard University Press.
Gertler, B. (2011). Self-knowledge and the transparency of belief. In A. Hatzimoysis (Ed.), *Self-knowledge* (pp. 125–145). Oxford: Oxford University Press.
Goldman, A. (2006). *Simulating minds*. Oxford: Oxford University Press.
Gordon, R. M. (2007). Ascent routines for propositional attitudes. *Synthese, 159*, 151–165.
Green, M., & Williams, J. (2007). Introduction. In M. Green & J. Williams (Eds.), *Moore's paradox: new essays on belief, rationality, and the first person* (pp. 3–36). Oxford: Oxford University Press.
'If only' (2020) Cambridge Dictionary [online]. Retrieved from: https://dictionary.cambridge.org/grammar/british-grammar/if-only. Accessed 20 June 2020.
Lawlor, K. (2009). Knowing what one wants. *Philosophy and Phenomenological Research, 79*, 47–75.
Lycan, W. (1996). *Consciousness and experience*. Cambridge: MIT Press.
Moore, G. E. (1942). A reply to my critics. In P. A. Schilpp (Ed.), *The philosophy of G.E. Moore* (pp. 535–677). Evanston: Tudor.
Moran, R. (2001). *Authority and estrangement*. Princeton: Princeton University Press.
Moran, R. (2004). Replies to heal, Reginster, Wilson, and Lear. *Philosophy and Phenomenological Research, 69*, 455–472.
Moran, R. (2012). Self-knowledge, "transparency", and the forms of activity. In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness* (pp. 211–236). Oxford: Oxford University Press.
Nichols, S., & Stich, S. (2003). *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*. Oxford: Oxford University Press.
Nozick, R. (1974). *Anarchy, state, and utopia*. Oxford: Basil Blackwell.
Paul, S. K. (2012). How we know what we intend. *Philosophical Studies, 161*, 327–346.
Saemi, A. (2015). Aiming at the good. *Canadian Journal of Philosophy, 45*, 197–219.
Scanlon, T. (1998). *What we owe to each other*. Oxford: Oxford University Press.
Schwitzgebel, E. (2010). Acting contrary to our professed beliefs, or the gulf between occurrent judgement and dispositional belief. *Pacific Philosophical Quarterly, 91*, 531–553.
Schroeder, T. (2004). *Three faces of desire*. New York: Oxford University Press.
Shoemaker, S. (1996). *The first-person perspective and other essays*. Cambridge: Cambridge University Press.
Silins, N. (2012). Judgement as a guide to belief. In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness* (pp. 295–327). Oxford: Oxford. University Press.
Sinhababu, N. (2017a). *Human nature: how desire explains action, thought, and feeling*. Oxford: Oxford University Press.
Sinhababu, N. (2017b). Desire and aesthetic pleasure. *Australasian Philosophical Review, 1*, 95–99.
Smith, M. (1994). *The moral problem*. New York: Oxford University Press.
Valaris, M. (2014). Self-knowledge and the phenomenological transparency of belief. *Philosophers' Imprint, 14*, 1–17.